BEYOND THE SCOPE: WILL HUMANOID ROBOTS DEFINE THE FUTURE OF ENDOSCOPY?

Yichen Wang[1], **Yuting Huang**[1], **Mario El Hayek**[2], **Vivek Kumbhari**[1], and **Michael B. Wallace**[1]

[1]Division of Gastroenterology and Hepatology, Mayo Clinic, Jacksonville, FL, USA

[2]Department of Medicine, Mayo Clinic, Jacksonville, FL, USA

Correspondence

Michael B. Wallace, MD, MPH, Mayo Clinic, 4500 San Pablo Rd, Jacksonville, FL 32224

For the past half-century, the technological trajectory of gastrointestinal endoscopy has been focused on visualization. Innovations in optics and high-definition sensors were designed to present a clearer picture to the human eye, placing the task of interpretation solely on the physician. Recent years have seen the integration of artificial intelligence (AI), in the form of computer-aided detection (CADe) and diagnosis (CADx).[1] Systems such as GI Genius have demonstrated the ability to reduce adenoma miss rates by acting as a "second observer"[2] However, these advancements remain fundamentally passive; they reside on the screen, disconnected from the physical activity of the procedure. The endoscopist still bears the total cognitive and physical load of navigating

the tortuous anatomy of the colon, creating a bottleneck where fatigue and variable motor skills contribute to significant inter-observer variability in procedural quality[3,4].

The discipline is now standing at the precipice of the epoch of embodied AI. This new paradigm represents the integration of advanced cognitive architectures, specifically vision-language-action (VLA) models, with physical robotic actuation. It marks the transition from systems that merely "see" to systems that "act." This shift is not merely incremental; it is a fundamental reimagining of what an endoscopic system can be. Embodied AI envisions a robotic agent capable of perceiving the complex, deformable environment of the gut, reasoning about anatomical landmarks in semantic terms, and autonomously executing kinematic maneuvers. This review analyzes the theoretical underpinnings of this convergence, exploring the specialized hardware of soft robotics and the emerging, disruptive potential of humanoid robots as autonomous operators.

The brain of an embodied agent requires capabilities far exceeding old convolutional neural networks. The field is adopting foundational vision-language models, to bridge medical knowledge and vision signals. Architectures like Endo-FM[5] and Endo-DINO[6] provide pre-trained foundational models for downstream tasks. The most significant leap is the Endoscopic Vision-Language-Action (EndoVLA) model.[7] EndoVLA is an end-to-end framework where a live video feed and a human text prompt generate direct control commands. This architecture solves the "semantic gap" in robotics by translating high-level intent, such as "resect that polyp," into low-level motor signals. To domesticate general-purpose VLMs for this safety-critical task, researchers employ a Dual-Phase Fine-Tuning strategy.[7] The model first undergoes Supervised Fine-Tuning to learn the "grammar" of robotic kinematics, followed by Reinforcement Fine-Tuning to hone precision through interaction.

A sophisticated cognitive engine is futile without a physical body capable of safe execution. One dominant approach to this challenge is the redesign of the instrument itself. Traditional semi-rigid endoscopes are prone to causing painful looping; consequently, the hardware for autonomous GI endoscopy is evolving toward soft robotics, prioritizing compliance and safety.[8,9] A paradigm-

shifting example is the spider-inspired magnetic soft robot developed at the University of Macau[10]. Modeled after the golden wheel spider, this device utilizes a magneto-active elastomer matrix to switch between rolling and climbing locomotion modes. Unlike passive capsules, this robot can actively climb over haustral folds and navigate upside down via magnetic anchoring. Simultaneously, Magnetic Capsule Endoscopy is maturing from teleoperated systems to autonomous platforms. Projects like MAGIC-AIM integrate VLM architectures with magnetic guidance, allowing the capsule to identify the pylorus and autonomously execute the traversal maneuver[7,11]. This automation ensures a complete map of the gastric mucosa is generated regardless of operator fatigue.

While soft robotics reinvent the tool, a parallel lineage proposes replacing the operator. The rise of general-purpose humanoid robots offers a tantalizing alternative: an autonomous agent that walks into the endoscopy suite, picks up the existing standard colonoscope, and performs the procedure using the same ergonomics designed for human hands.

Backward compatibility drives this approach. Hospitals worldwide have billions of dollars invested in standard flexible endoscopes. Soft robotic solutions require abandoning this infrastructure for expensive, proprietary magnetic guidance systems. A humanoid robot, conversely, is a "drop-in" replacement. It utilizes the current capital equipment, navigating the colonoscope via the control wheels and shaft torque just as a human fellow would. Recent feasibility studies have demonstrated that humanoids can effectively manipulate laparoscopic tools via teleoperation, suggesting that the transition to flexible endoscopy is a matter of software refinement rather than hardware impossibility[12].

The cold hard truth is we are stalled by clinical and technical inertia, primarily the "Dexterity Gap." We have spent decades refining the endoscope as a tool for humans, yet we now face a reality where our current hardware is a legacy anchor. The industry is hooked on high-margin, semi-disposable scopes that are fundamentally incompatible with the precision of high-torque robotic motors[9] If we continue to ignore the friction between 20th-century ergonomics and 21st-century AI, we risk creating a perpetual pilot phase that never reaches the patient's bedside. Hospitals

are loath to cannibalize their existing infrastructure, yet they complain about a shortage of hands to hold the scope. The human hand is a marvel of compliance and sensory feedback, capable of sensing the subtle resistance of a loop forming in the sigmoid colon through the shaft of the scope. Current commercial humanoid hands are often rigid, under-actuated grippers designed for gross manipulation, not the delicate interplay of torque and translation required in endoscopy[13]. For a humanoid to be viable, it requires reliable, dexterous hands equipped with high-fidelity tactile sensors at a low price point. The robot must be able to manipulate the angulation wheels with its thumb while simultaneously torquing the shaft with its right hand, a bimanual coordination task that remains an open challenge in robotics control.

Furthermore, there is a critical scarcity of training data known as the "proprioceptive void." While we have millions of hours of endoscopic video, we have virtually zero data recording the actions of the endoscopist's hands synchronized with that video[2,9]. We know what the colon looks like, but we do not know what the doctor's hands were doing to produce that view. To train a humanoid EndoVLA, we must bridge this gap by instrumenting expert endoscopists with motion-capture gloves or sensors during live procedures, creating massive "Vision-Action" datasets. Once this data exists, the potential for scaling is boundless. A model trained on the collective motor skills of the world's best endoscopists could be deployed to a humanoid in a rural clinic, instantly granting it the dexterity of a master clinician.

The humanoid approach warrants optimism. General robotics outpaces specialized medical tools due to massive tech investment. As costs fall, humanoids will address the global gastroenterologist shortage. A future where robots perform screening colonoscopies 24/7 is a converging probability. Economic incentives are powerful. As reimbursement fluctuates, a robotic workforce becomes an existential necessity. Continuous operation eliminates the 'off-hours' penalty where outcomes suffer due to staffing. A robot does not sleep, and its hand never shakes at the end of a shift. This consistency is the true promise of the revolution.

The integration of the brain and body occurs in the navigation loop. FM and VLM architectures

provide the reasoning required for autonomous tracking through language-conditioned tasks. By processing a prompt and video feed to output discrete motion commands, these systems offer Zero-Shot Generalization, allowing the tracking of novel objects like surgical clips without retraining[7].

Generative AI necessitates rigorous safety. Researchers are employing Safe Reinforcement Learning using Control Barrier Functions that act as mathematical virtual walls, analyzing actions against safety constraints to prevent perforation.[14] To bridge the Sim-to-Real gap, agents are trained in high-fidelity simulators using domain randomization, ensuring robust performance despite the chaotic textures of a living patient.

The convergence of embodied AI and GI endoscopy signifies a fundamental restructuring of gastroenterological care. The technological pillars are in place: Vision-Language Models have solved the semantic gap, enabling robots to understand medical intent; and two distinct physical lineages are vying for dominance. On one hand, soft robotics offers a safety-first reinvention of the instrument; on the other, humanoid robotics offers a scalable, backward-compatible reinvention of the operator. As we look toward the next decade, the standard of care is poised to shift from manual operation to supervisory control. Future screening protocols may involve patients being treated by a humanoid agent that wields standard instruments with the precision of a machine and the adaptability of a human. This evolution promises to democratize high-quality care, reducing the physical burden on physicians and ensuring that every patient receives a procedure defined by the collective intelligence of the field rather than the variability of a single operator. The era of the passive tool is ending; the era of the embodied partner has begun.

For practicing endoscopists, a "robot-ready room" can log wheel rotations, torque, and finger kinematics during routine lists to generate vision–action corpora without slowing workflow. Health systems absorbing the screening-age drop to 45 years could designate supervised data-collection blocks and trial overnight humanoid-run screening lines with remote oversight to clear backlogs[15]. These experiments demand audit trails, explainable policies, and governance that preserve human accountability even as robots offload repetitive motor work[16].

Keywords: Artificial Intelligence, Humanoid Robots, Soft Robotics, Vision-Language Models

ABBREVIATIONS

AI, Artificial Intelligence; CADe, Computer-Aided Detection; CADx, Computer-Aided Diagnosis; EndoVLA, Endoscopic Vision-Language-Action; FM, Foundation Model; GI, Gastrointestinal; VLM, Vision-Language Model.

References

References

1. Chan HP, Samala RK, Hadjiiski LM, and Sahiner C. Computer-aided diagnosis in the era of deep learning. Medical Physics 2020;47:e218–e227.

2. Wallace MB, Sharma P, Bhandari P, et al. Impact of Artificial Intelligence on Miss Rate of Colorectal Neoplasia. Gastroenterology 2022;163:295–304.

3. Chan MY, Cohen H, and Spiegel BM. Fewer polyps detected by colonoscopy as the day progresses at a Veteran's Administration teaching hospital. Clinical Gastroenterology and Hepatology 2009;7:1217–23.

4. Kaminski MF, Regula J, Kraszewska E, et al. Quality indicators for colonoscopy and the risk of interval cancer. New England Journal of Medicine 2010;362:1795–803.

5. Wang Y, Zhang Z, Liu H, et al. Endo-FM: Foundation Model for Endoscopy Video Analysis via Large-scale Pre-training. arXiv preprint arXiv:2306.16741. 2023. URL: https://arxiv.org/abs/2306.16741.

6. Liu Y, Wang Y, Zhang Z, Chen H, and Wu J. Endo-DINO: Self-Supervised Pre-training for Endoscopy with Vision Transformers. arXiv preprint arXiv:2501.05488. 2025. URL: https://arxiv.org/abs/2501.05488.

7. Wang R, Hong Y, He Y, Wu J, and Yuan W. EndoVLA: Progressive Endoscopic Vision-Language-Action Model. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer, 2024:446–56. DOI: 10.1007/978-3-031-72083-3_42. URL: https://doi.org/10.1007/978-3-031-72083-3_42.

8. Shah SG, Brooker JC, Thapar C, Williams CB, and Saunders BP. Patient pain during colonoscopy: an analysis using real-time magnetic endoscope imaging. Endoscopy 2002;34:435–40.

9. Martin JW, Scaglioni B, Norton JC, et al. Enabling the future of colonoscopy with intelligent and autonomous magnetic manipulation. Nature Machine Intelligence 2020;2:595–602.

10. Wang Z, Shi K, and Xu Q. A Millimeter-Scale Magnetic Soft Robot With Spider-Inspired Multi-Modal Locomotion and Object Manipulation. IEEE/ASME Transactions on Mechatronics 2022;27:4676–87.

11. Hong Y, Zhang Y, and Meng MQH. Autonomous pylorus traversal in magnetic capsule endoscopy via deep reinforcement learning. IEEE Transactions on Robotics 2021;37:1094–108.

12. Wang R, Hong Y, He Y, Wu J, and Yuan W. A Feasibility Study on Humanoid Robot Teleoperation for Laparoscopic Surgery. arXiv preprint. 2024.

13. Piazza C, Grioli G, Catalano MG, and Bicchi A. A Century of Robotic Hands. Annual Review of Control, Robotics, and Autonomous Systems 2019;2:1–32.

14. Hong Y, Wang R, He Y, Wu J, and Yuan W. Safe Reinforcement Learning for Autonomous Flexible Endoscopy. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023:9460–6. DOI: 10.1109/ICRA48891.2023.10161476. URL: https://doi.org/10.1109/ICRA48891.2023.10161476.

15. Crockett SD and Ladabaum U. Potential Effects of Lowering Colorectal Cancer Screening Age to 45 Years on Colonoscopy Demand, Case Mix, and Adenoma Detection Rate. Gastroenterology 2022;162:984–986.e5.

16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine 2019;25:44–56.